Correspondence as the Primary Measure of Quality for Web Archives: A Grounded Theory Study

Brenda Reyes Ayala[1][0000-0002-9342-3832]

University of Alberta, Edmonton AB T6G 0T5, Canada brenda.reyes@ualberta.ca

Abstract. Creating an archived website that is as close as possible to the original, live website remains one of the most difficult challenges in the field of web archiving. Failing to adequately capture a website might mean an incomplete historical record or, worse, no evidence that the site ever even existed. This paper presents a grounded theory of quality for web archives created using data from web archivists. In order to achieve this, I analysed support tickets submitted by clients of the Internet Archive's Archive-It (AIT), a subscription-based web archiving service that helps organisations build and manage their own web archives. Overall, 305 tickets were analysed, comprising 2544 interactions. The resulting theory is comprised of three dimensions of quality in a web archive: correspondence, relevance, and archivability. The dimension of correspondence, defined as the degree of similarity or resemblance between the original website and the archived website, is the most important facet of quality in web archives, and it is the main focus of this work. This paper's contribution is that it presents the first theory created specifically for web archives and lays the groundwork for future theoretical developments in the field. Furthermore, the theory is human-centred and grounded in how users and creators of web archives perceive their quality. By clarifying the notion of quality in a web archive, this research will be of benefit to web archivists and cultural heritage institutions.

Keywords: web archiving · information quality · quality assurance · grounded theory.

1 Introduction

In 1996, the Internet Archive was founded in San Francisco with the goal of building a universally accessible digital library. The Internet Archive began using a web crawler to periodically take snapshots of websites and store them as historical records. Internet users could then access these archived websites using the Wayback Machine, a special piece of software developed by the Internet Archive. As the World Wide Web evolved, the pace at which websites changed their content and appearance accelerated dramatically: websites were redesigned or disappeared altogether, additional materials such as video and audio were added, and social media began to emerge. Often the Internet Archive's cache was the only record of how a website had evolved or that it had existed at all. By the dawn of the new millennium, the practice of “web archiving,” as it became known, had spread beyond the Internet Archive. Organisations such as national libraries, governments, and universities began also to archive websites for the purpose of preserving their digital heritage.

Though enormous strides have been made, web archiving today remains a complicated and technically-challenging endeavour. New web technologies emerge constantly, and web archivists struggle to keep up. Creating an archived website that is as close as possible to the original, live website remains one of the most difficult challenges in the field. Failing to adequately capture a website might mean an incomplete historical record or, worse, no evidence that the site ever even existed. It is in the context of these challenges that this research takes place.

In the field of web archiving, there have been few comprehensive definitions of quality. One such definition was put forward by Masanès [13]. He defined quality in a web archive as having the following characteristics:

1. the completeness of material (linked files) archived within a target perimeter
2. the ability to render the original form of the site, particularly regarding navigation and interaction with the user [13]

This definition of quality, though useful, is centred on the technological tools needed to archive websites. Terms such as “target perimeter” refer to the configuration of web crawlers. If the web archive was created using alternative methods, or if crawlers were replaced in the future by newer, more efficient tools, then Masanés’ definition would become obsolete. Another problem is that it lacks a human element; one never finds out what quality might mean to the users and creators of web archives. This definition ignores the context in which a web archive exists and whether or not it meets the needs of its users. A more robust definition of quality in web archives is needed, one that is independent of the technology currently in use to create web archives and that incorporates a human element. The lack of proper definitions of quality is indicative of a larger problem in the field of web archiving. The technical developments in the field have far outpaced the development of proper theoretical tools or models. Over two decades into its history, web archiving still lacks a theoretical underpinning. Essentially, we have technological tools to build web archives, but no conceptual tools to understand them.

The goal of this research is to build a theory of quality for web archives that is grounded in user-centred data. This goal leads to the following research question: What is the human-centred definition of quality for web archives? This paper presents the first theory created specifically for web archives and lays the groundwork for future theoretical developments in the field. Furthermore, the theory is human-centred and grounded in how users and creators of web archives perceive quality. It also marks the first application of grounded theory to the discipline of web archiving. By clarifying the notion of quality in a web archive, this research will be of benefit to web archivists and cultural heritage institutions who seek to improve the Quality Assurance processes for their organisations.

2 Previous Work

Over the last decade, researchers have begun to study the topic of quality for web archives. Some have also attempted to operationalize individual aspects of quality and to create metrics to effectively measure it. In their paper, Spaniol et al. (2009)[17] are primarily concerned with the quality of a crawl, not with replay of the archived website itself. The authors introduce the concept of (temporal) coherence for a web archive. The contents of a web archive are considered to be coherent if they appear to be "as of" time point x or interval *[x;y]*. In a web archive, coherence defects can occur during the crawl, a process which can take anywhere from a few minutes to weeks for large websites. The authors explored ways to visualize coherence defects in a web archive, so that crawl engineers could detect them and adjust their crawling strategies accordingly.

In a later paper, Denev, Mazeika, Spaniol, and Weikum [7] introduced the Sharp Archiving of Website Captures (SHARC) framework for data quality in web archiving. This framework included two measures of data quality for capturing websites: *blur* and *coherence*. Blur was defined as the expected number of page changes that a time-travel access to a site capture would accidentally see, instead of the ideal view of a instantaneously captured, "sharp" site. This value needed to be minimized in order to achieve a high-quality capture. The authors defined coherence as the number of unchanged and thus coherently captured pages in a site snapshot. Coherence needed to be maximized in order to achieve a high-quality capture.

The work of Ainsworth, Nelson, and Van de Sompel [2] further expanded the notion of temporal coherence in a web archive. They pointed out that archived web pages are composite objects and that, because of the constantly changing nature of the web, many elements and pages from the archived website will have been collected before or after the date presented by the Wayback Machine. The final, archived website presented to the user is often a patchwork collection of HTML pages, images, and scripts from different dates and is thus temporally incoherent. They defined the *temporal coherence* of an archived website (which they call a memento) in the following way: "an embedded memento [is] temporally coherent with respect to a root memento when it can be shown that the embedded memento's representation existed at the time the root memento was captured". Ainsworth, Nelson, and Van de Sompel [2] also specified an extension of their defined coherence states that involved calculating the similarity, or lack thereof, between two archived versions of the same website (or, as the authors put it, between two mementos). This comparison, which they called a "content pattern", takes into account not just the time of archival, but also the content of the two mementos in order to determine coherence. It is important to note that according to the authors, the additional computational cost of calculating these comparisons "may render content patterns unsuitable for casual archive use or in restricted bandwidth conditions".

Ainsworth and Nelson [1] were also concerned with defining quality as meeting measurable characteristics. Their work elaborates on the notion of coherence put forward by Denev, Mazeika, Spaniol, and Weikum [7]. They equate the completeness of a web archive to its coverage; in other words, a complete web archive does not have undesired or undocumented gaps.

Other researchers have addressed the notion of completeness in a web archive. Web archives do not contain complete and perfectly accurate copies of every single website they intend to capture; the dynamic nature of the web makes this almost technically impossible. However, not all missing elements are created equal. Many archived websites are missing elements but still retain most of their intellectual content, while other archived websites, such as maps, are rendered unusable due to missing elements. Brunelle, Kelly, SalahEldeen, Weigle, and Nelson [6] made precisely this point when they examined the importance of missing elements or resources and their impact on the quality of archived websites in their paper.

When deploying crawlers to capture a website, some crawl engineers pay special attention to embedded resources. Embedded resources are files, such as images, videos, or CSS stylesheets, that are present and referenced in a website. A user might not notice their presence, but embedded resources play a key role in ensuring the website looks and operates in the correct way. To this end, crawl engineers might calculate a percentage of missing embedded resources M_m in an archived website and use it to estimate the overall quality of the site. Brunelle, Kelly, SalahEldeen, Weigle, and Nelson [6] showed that M_m is not always consistent with human judgments of the quality of an archived website and was thus not a suitable metric for measuring the "damage" to an archived website caused by missing embedded resources. Instead, the authors proposed a new metric to assess this damage that is based on three factors: the MIME type, size, and location of the embedded resource [6].

AlNoamany, Weigle, and Nelson [3] also addressed quality problems that could affect the coherence of a web archive, such as off-topic web pages. Many web archives are topic-specific: they collect and preserve websites that cover a single topic or news event, such as Human Rights or the Arab Spring of 2010. Off-topic web pages are defined as those that have, over time, moved away from the initial scope of the page. This can occur because the page has been hacked, its domain has expired, or the service has been discontinued. The authors compiled three different Archive-It collections and experimented with several methods of detecting these off-topic webpages and with how to define threshold that separates the on-topic from the off-topic pages. According to their results, the cosine similarity method proved the best at detecting off-topic web pages. The authors also experimented with combining several similarity measures in an attempt to increase performance. The combination of the cosine similarity and word count methods yielded the best results, with an accuracy equal to 0.987, $F = 0.906$, and $AUC = 0.968$ [3].

Banos et al. [5] introduced the concept of website archivability, defined as the "sum of the attributes that make a website amenable to being archived". The more easily it was to archive a website, the greater its archivability. The authors introduced the CLEAR+ method to determine the archivability of a website. According to CLEAR+, an archivable website is accessible (a web crawler can traverse it easily); complies to common, accepted technical standards; cohesive (its components are not dispersed across different locations on the web); and uses descriptive metadata.

In their work, Poursardar and Shipman [15] conducted a user study to explore how users view the boundaries of web resources in institutional web archives, especially as compared to personal archives. Participants were recruited through Amazon's Mechanical Turk and presented with pairs of main/primary web pages. The authors found that, when accessing institutional web archives, users expect the main content to be preserved, as well as additional linked content, advertisements, and author information. In other words, users who access institutional web archives have expectations as to what content should be preserved that are similar to users accessing personal archives.

Kiesel, Kneist, Alshomary, Stein, Hagen, and Potthast [11]paper focused on the reproduction quality of archived websites. To this end, they introduced the Webis Web Archiver tool, which relied on emulating user interactions with a web page while recording all network traffic. In order to evaluate their tools, the researchers recruited human evaluators (recruited through Amazon's Mechanical Turk) to assess web pages in their dataset. The authors defined reproduction quality as thus: "the more individual users that scroll down a web page are affected in their perception or use of the web page by visual differences between the original web page and its reproduction, the smaller the reproduction quality for that web page." Reproduction quality was assessed on a 5-point Likert scale to account for different levels of perceived severity, ranked from no effect (score 1) to unusable reproduction (score 5). Some examples of the assessment scale used were:

- Score 1 (not affecting): Parts of the page are moved up and down a bit.
- Score 2 (small effect on a few visitors): Social media buttons, ads, or unimportant images or text are missing.
- Score 3 (small effect on many or all visitors): Comments on the main content are missing.
- Score 4 (affects, but page can still be used): Striking difference in colour, background, or layout.
- Score 5 (unusable page): Important/main content is missing and/or visitors can't use the right page due to differences.

As Kiesel et al.'s work acknowledges [11], many quality problems arise as a result of the replay process because current technologies such as the Wayback Machine are unable to adequately render the archived website as it originally appeared. The lack of adequate technologies to address quality problems in web archives was highlighted by Klein et al. [12] in their 2019 paper. The authors stated that current web archiving technologies were optimized to either: 1) operate at scale or 2) provide high-quality archival captures, but not both. To address this imbalance, they introduced the Memento Tracer framework, which aimed to achieve both quality and quantity, by allowing the curator to determine the desired components of a web resource that should be archived. Klein et al.

acknowledged that quality in web archives is often subjective, and thus focused on the extent to which URIs that should be captured are actually captured. The authors "expect that a high-quality archival record to contain at least the same number of URIs as its live website version" [12]. In other words, a high-quality web archive is *complete*.

The recent focus on the issue of quality in web archives is significant and has resulted in a better understanding of what constitutes a high-quality archived website, and contributed to the emergence of tools designed to improve quality. However, these approaches have been somewhat piecemeal; some researchers focus on completeness, others on coherence, others on relevance, etc. Comprehensive notions of quality are still forthcoming. It is also important to note that assessments of quality obtained from Mechanical Turk users might differ from assessments of quality obtained from web archivists, who are experienced in the processes of creating web archives and might have different or higher standards for preserving web content due to institutional goals and mandates. The research presented here aims to address some of these gaps in the literature.

3 Methodology: Building a Theory of Quality in a Web Archive

In the 60s, Barney Glaser and Anselm Strauss created the methodology of Grounded Theory (GT), which they defined as "the discovery of theory from data - systematically obtained and analysed in social research" [8]. For the authors, theory was not a perfected product that explains all facets of a phenomenon, but a process, an ever-developing entity. GT is an inductive methodology; working closely from the data, the researcher begins the work of generating a theory. GT is optimal for this research problem for the following reasons:

1. There are no existing models or theories in the area of web archiving. GT is appropriate for situations such as these where a field is relatively unexplored and there is a need for theoretical explanations and models [10].
2. GT is user-centred. As its name implies, GT is heavily "grounded" in rich contextual data gathered from empirical research with actual persons [10].
3. GT is iterative. GT research involves the *constant comparison* method, which has the researcher constantly compare the emerging model/theory to the data. This allows the researcher to continually redefine a model and to become aware when no new information is emerging.

3.1 Data Gathering and Processing

The Internet Archive's Archive-It (AIT) is a subscription-based web archiving service that helps organisations build and manage their own web archives. Archive-It is currently the most popular web archiving service, with over 600 clients (called "partners") consisting of universities, state libraries and archives, museums, and national libraries in several countries [4]. The accounts of Archive-It clients are managed by a team of partner specialists. When a client encounters a problem with Archive-It, she first opens a support ticket using Zendesk, a popular customer-service platform. The ticket is received by a partner specialist, who is then responsible for addressing the issue. These initial tickets are part of the "Level 1" support. If the partner specialist determines that a problem is more serious or highly-technical in nature, the issue becomes a "Level 2" and a ticket is opened in JIRA, another issue-tracking platform. There is one support engineer who is responsible for addressing these Level 2 tickets. If he determines that the problem requires more extensive technical efforts, he will convert it to a "Level 3" ticket, which is then addressed by the software engineers at the Internet Archive.

AIT support tickets are a rich source of information regarding quality problems in web archives. They contain the opinions and views of individuals who are experienced creators of web archives, well-versed in web archiving processes, and familiar with institutional web preservation goals, whether they be clients or the partner specialists themselves. They contain rich descriptions of how quality problems are detected, analysed, and addressed, and are thus an ideal dataset for studying quality in all its dimensions.

The first step was to obtain Archive-It support tickets in order to analyse them. Since these tickets belonged to the Internet Archive, I negotiated a researcher agreement with the organisation to obtain support tickets from the years 2012 through 2016. The tickets received comprised a wide variety of institutions reflecting AIT's client base, from national libraries, to private organisations, to universities and museums from Europe, North America, and Asia. After the tickets were cleaned, I randomly selected the same amount of tickets for each year from 2013 through 2016. This randomisation approach was taken to minimise the selection bias that might have occurred if I had manually chosen which tickets to analyse. The final dataset of 645 tickets was then imported into the NVivo software package, a popular program for performing qualitative data analysis [16].

Among other conditions, the research agreement stipulates that the researcher anonymise any personal or institutional information present in the tickets, as well as any other potentially identifying information. In order to comply with the terms of this agreement, all the information presented in this paper has been anonymised: identifying elements such as personal names, names of institutions, and website addresses have been removed or changed.

Data Analysis The tickets collected were Level 1 support tickets that had been submitted by AIT client. They included the initial question submitted by the client, the response given by the AIT partner specialist, and any subsequent communication between the two. As has been previously noted, Level 2 and Level 3 support tickets represent communication between the AIT support engineer and the team of software engineers. Because these tickets do not involve the AIT clients and are highly technical in nature, they do not contain the opinions of users and creators of web archives. Therefore, they were not considered relevant to the project and were not requested.

It is important to note that not all the AIT tickets deal with issues of quality in a web archive. Quite a few deal with collection management issues, such as how to manage user accounts for a collection of web archives, storage limitations, and questions about the privacy or public access to archived content. This research focuses on tickets in which the client discusses a perceived flaw in an individual archived website or an entire web archive. From prior experiences, I had seen that these types of tickets are the most likely to deal with issues of quality.

Support tickets not pertaining to quality issues were classified as such and separated from the main data of interest. Each ticket analysed consisted of the original ticket submitted by the client, the response sent by the AIT partner specialist, and any subsequent interactions between them. Tickets could be quite brief, consisting of three interactions (the original client ticket, the partner specialist's response, and the client's response), or they could have many interactions over time, spanning weeks or even months. Table 1 lists the number of tickets and interactions about quality that were analysed, which totalled 305 tickets and 2544 interactions.

Table 1. Number of Tickets and Interactions About Quality Analysed Per Year

Year	No. tickets about quality analysed	No. interactions analysed
2012	74	478
2013	65	492
2014	67	540
2015	58	528
2016	41	506
Total	305	2544

These support tickets were analysed using the GT techniques of open coding and theoretical memos to identify the main concepts and categories present in the data. According to the precepts of GT, after several rounds of coding, the researcher will reach *saturation*, a state when nothing new is being extracted from the data. Per the guidelines of Grounded Theory, only the core categories (that is, the ones that explain most of the variation in quality) are part of the final theory. In order to increase the quality and rigour of the study, I engaged in purposeful peer review. University professors were periodically invited to audit the entire research project, including the codebook, preliminary findings, and core categories. In addition to peers, employees of the Internet Archive were also invited to see the findings and comment on them.

4 Findings and Discussion

4.1 Core Categories

The grounded theory presented here consists of three dimensions (or core categories) that determine the quality of a web archive: correspondence, relevance, and archivability.

1. Correspondence: degree of similarity, or resemblance, between the original website and the archived website
   (a) Visual correspondence: similarity in appearance between the original website and the archived website
   (b) Interactional correspondence: the degree to which a user's interaction with the archived website is similar to that of the original
   (c) Completeness: the degree to which the archived website contains all of the components of the original
2. Relevance: pertinence of the contents of an archived website to the original website
   (a) Topic relevance: degree to which an archived website (or a web archive) includes only content that is closely related to that of the original website or the topic of the larger web archive
   (b) Size relevance: the similarity in size of the archived website to the original website
3. Archivability: degree to which the intrinsic properties of a website make it easier or more difficult to archive

Taken together, these three dimensions meet the requirements specified by Barney Glaser [9]. As core categories, they account for most of the behaviour of web archivists towards the quality of web archives that was seen in the data. Of all the three core categories examined, the dimension of correspondence was by far the most important, with 852 mentions across 226 tickets, much more than relevance (451 mentions across 127 tickets) and archivability (101 mentions across 78 tickets). **Due to its importance, the dimension of correspondence is the main focus of this work.**

4.2 Visual Correspondence

When describing a quality problem in the tickets, AIT clients will often compare the archived website to the original website. AIT clients have a strong idea of what the archived website should look or behave like and are quick to report any discrepancies. Table 2 displays some examples of problems with visual correspondence. In these, AIT clients point out how the visual appearance of the archived website does not match that of the original. Clients express these comparisons in a number of ways. One way is by including a direct link to the original website in their tickets. This allows the partner specialist to make quick comparisons between the live site and the archived website and note the differences. Table 2 shows some examples of tickets where the clients made these explicit comparisons. In ticket 103, the client tells the AIT partner specialist to check the live website for the "proper" version ("how it should look"). Many more tickets do not include the URL for the original website, but still explicitly compare it to the archived version. Some of these instances are also shown in Table 2. The clients describe the archived website as being problematic: it is "a bit off" (ticket 36), it "does not display properly" (ticket 302), or does not capture the "the look and feel" of the original (ticket 3420).

Table 2. Examples of Problems with Visual Correspondence

Ticket Name	**Text of the Ticket**
ticket 103	I have done a crawl of the following: http://www.___.org/remembering/ and the YouTube video display is problematic in Wayback on the pages. While the host report has the YouTube videos captured, they are not showing up on the web pages. See http://wayback.archive-it.org/yyhttp://www.___.org/remembering/life-work http://www.___.org/remembering/life-work for how it should look.
ticket 260	On the new http://www.stateu.edu/academics page we are not capturing the background images. I cannot figure out why since we are capturing other images from the same directory
ticket 33	(see http://___.uk/roman-scrolls compared to http://wayback.archive-it.org/http://___.uk/roman-scrolls) Poets - Text next to the portraits should change as you scroll over the navigation bar. (http://___.uk/ vs http://wayback.archive-it.org/http://poetry.___.uk/)
ticket 36	I also noticed that the display for your www.nzlibrary.edu pages was a bit off.
ticket 302	We're having some trouble with our Facebook site captures not displaying properly (or at all, really).
ticket 3420	One thing related though, the page is not capturing its look and feel well...Any suggestions? It's missing the background and objects are not in the right locations.

Table 3. Examples of Problems with Interactional Correspondence

Ticket Name	**Text of the Ticket**
ticket 114	The site renders fine and you can hover over the progress bar for the videos and see that the frames are captured, but the video won't play.
ticket 27	Clicking "View all comments" under an update does not reveal the comments.
ticket 33	the interactive floorplan isn't working as it should do - the text should appear over the map when you click on it, rather than in a list underneath.
ticket 3276	I know I've captured the video but it doesn't play on the web page
ticket 3284	When i click on it, it briefly flashes to the homepage and then it displays a URL with the nationalscience URL in it twice.
ticket 74	In some cases I hear audio but see no video

Table 4. Examples of Problems with Completeness

Ticket Name	Text of the Ticket
ticket 114	It looks like what is happening is that the video files themselves have not been captured
ticket 33	there should be a Google search bar at the top of both websites.
ticket 296	on all most every blog that we have captured from blogspot the Wayback Machine does not include the subsequent pages beyond the first.
ticket 311	We're still having some trouble capturing the JavaScript menu at the top of the main page. I know that JS can be wonky.
ticket 3117	The News pages (which are located under each individual sport) are being captured, but the actual articles that are listed and linked out are not.
ticket 74	The issue with this seed is that for all previous crawls we were able to capture main text for individual ___ articles, but not comments.

4.3 Interactional Correspondence

Interactional correspondence was defined as a sub-category of the correspondence dimension of quality. A problem with interactional correspondence occurs when a user's interaction with the archived website is different from that of the original, unexpected, or deficient. For example, on the live website, a web archivist clicks on a link and is taken to the corresponding target of that link, that is, another webpage. She expects the same thing to happen on the archived version of the original page. If it does not, and she is not taken to a different webpage, the archived website lacks interactional correspondence. Problems with interactional correspondence occur when there is a mismatch between a user's expectation of website behaviour and the actual behaviour displayed by the archived website.

Similarly, examples of problems with interactional correspondence are shown in Table 3. When the clients attempt to interact with the archived website as they would with the original, they report unexpected behaviours: the text in the interactive floor plans does not display in the correct location (ticket 33), a page displays only very briefly and then redirects to another location (ticket 3284), and clicking on a button does not display the comments (ticket 27). Video content in web archives is also difficult to replay (tickets 114, 3276, and 74).

It is important to note that these codes are not independent of each other. It is common for a low-quality archived website to have many problems, from missing pages to unexpected behaviours Some quality problems straddle several categories. For example, ticket 260 from Table 2 is presented as an instance of a visual correspondence problem, since the archived site does not include the background images as the original does. However, the same ticket can also be classified as a completeness problem, since the site is missing images (intellectual content) that it should contain. In fact, many (though not all) archived websites that exhibit mismatched appearance and behaviours do so because they are missing important files that provide needed visual elements or functionality. Though the categories are separate, they are often linked.

4.4 Completeness as a Type of Correspondence

Completeness has already been described as the completeness of an archived website as it relates to the original. A perfectly complete archived website contains all of the components of the original. A completeness problem occurs when the original website's content has not been captured or is not present in the archive. Lack of completeness is caused by the absence of needed content. Table 4 displays examples of completeness problems, where the clients note that an archived website is missing content that assumed to be present in the original. They report missing search boxes (ticket 33), articles (ticket 3117), menus (ticket 311), videos (ticket 114), comments (ticket 74), and in some cases, even archived websites that are missing many pages (ticket 296).

In the literature that was reviewed, completeness is often seen as a major aspect of quality, sometimes even equated with quality itself. It is present in the work of Masanés [13], Ainsworth and Nelson [1], Brunelle et al. [6], and Klein et al. [12]. It is therefore tempting to see completeness as its own separate dimension of quality in web archives, different from correspondence; however, this is a fallacy. An archived website can have a lack of correspondence with the original website yet still be perfectly complete. For example, it can have all the same components of the original, yet still look or behave differently from it. However, the reverse is not true: an archived website cannot be incomplete, yet still have 100% correspondence with the original. In logic, correspondence is known as a *necessary cause*: "If x is a necessary cause of y, then the presence of y necessarily implies the presence of x with a probability of 100%. The presence of x, however, does not imply that y will occur." [14]. The presence of a lack of completeness (y) always implies the presence of a lack of correspondence (x); however, the presence of correspondence does not imply a lack of completeness. Therefore, completeness is not a core category in the theory, but rather a subcategory.

The work presented in this paper is delimited because it is specific to small or medium-size web archives that are focused on covering a single topic or an event. It is not meant to describe larger web archives such as the *.gov* or *.fr*, which preserve an entire country's national domain. The theory of quality in web archives presented here makes an important assumption: that there exists a live version of a website to which the archived version can be compared. However, the correspondence of an archived webpage might not always be easily known. For example, if the original site has been lost, there is no way to compare it to the archived version, so a measure of correspondence cannot be calculated.

5 Conclusion and Future Work

This paper makes the following contributions:

1. The paper presents the first application of grounded theory to the discipline of web archiving.

2. It introduces the first theory of quality developed specifically about web archives, and lays the groundwork for future theoretical and practical developments in the field.
3. The theory is human-centred and grounded in how subject-matter experts in the field of web archiving perceive quality.
4. The theory is *comprehensive*, incorporating and unifying the work of previous researchers on web archives.
5. The theory is independent of the technology currently in use to create web archives, making it suitable to a wide variety of platforms, preservation contexts, and situations.

Taken together, the theory presented here represents the majority of quality problems seen in topic-centred or event-driven web archives today. According to Glaser and Strauss, a grounded theory must closely fit the data and also be clear and flexible [8]. This last requirement is especially important. A theory must be flexible enough that a user who applies the theory is able to adjust it and reformulate it as she encounters new data and situations. For example, if in the future, new technologies were developed to capture dynamic web content more successfully, the notions of visual correspondence, interactional correspondence, and completeness would still be relevant to quality in web archives. As Glaser and Strauss state "evidence and testing never destroy a theory (of any generality), they only modify it. A theory's only replacement is a better theory" [8].

Having clear concepts based on experts perceive the issue of quality can lead to the successful creation of metrics, methods, and tools that will enable web archivists to measure the quality of their web archives. For example, in order to measure the correspondence of a web archive, a program could be developed that would navigate to both the live website and its archived counterpart, and then calculate some measure of similarity between them in terms of visual correspondence, interactional correspondence, and completeness. Once the software to measure correspondence has been built, experiments could be carried out to determine which metrics perform best. Details such as these would need time and effort to be adequately worked out, but the results would ultimately lead to higher quality web archives, and thus, a better and more complete historical record.

References

1. Ainsworth, S.G., Nelson, M.L.: Evaluating sliding and sticky target policies by measuring temporal drift in acyclic walks through a web archive. International Journal on Digital Libraries **16**(2), 129–144 (2015). https://doi.org/10.1007/s00799-014-0120-4
2. Ainsworth, S.G., Nelson, M.L., Van de Sompel, H.: A framework for evaluation of composite memento temporal coherence. Computing Research Respository (CoRR) **abs/1402.0928** (2014), http://arxiv.org/abs/1402.0928
3. AlNoamany, Y., Weigle, M.C., Nelson, M.L.: Detecting Off-Topic Pages in Web Archives, vol. 9316, pp. 225–237. Springer International Publishing, Cham, Switzerland (2015)

4. Archive-It: Learn more (2020), https://archive-it.org/learn-more
5. Banos, V., Manolopoulos, Y.: A quantitative approach to evaluate website archivability using the CLEAR+ method. International Journal on Digital Libraries pp. 1–23 (2015). https://doi.org/10.1007/s00799-015-0144-4
6. Brunelle, J., Kelly, M., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: Not all mementos are created equal: measuring the impact of missing resources. International Journal on Digital Libraries pp. 1–19 (2015). https://doi.org/10.1007/s00799-015-0150-6
7. Denev, D., Mazeika, A., Spaniol, M., Weikum, G.: The SHARC framework for data quality in web archiving. The VLDB Journal **20**(2), 183–207 (Mar 2011). https://doi.org/10.1007/s00778-011-0219-9
8. Glaser, B., Strauss, A.: The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine Transaction (2009), http://amazon.com/o/ASIN/0202302601/
9. Glaser, B.: Theoretical Sensitivity: Advances in the Methodology of Grounded Theory. The Sociology Press, Mill Valley, CA (1978)
10. Grbich, C.: Qualitative Data Analysis: An Introduction. SAGE Publications Ltd, London, 2nd edn. (11 2012)
11. Kiesel, J., Kneist, F., Alshomary, M., Stein, B., Hagen, M., Potthast, M.: Reproducible web corpora: Interactive archiving with automatic quality assessment. Journal of Data and Information Quality **10**(4) (Oct 2018). https://doi.org/10.1145/3239574, https://doi.org/10.1145/3239574
12. Klein, M., Shankar, H., Balakireva, L., Van de Sompel, H.: The memento tracer framework: Balancing quality and scalability for web archiving. In: Doucet, A., Isaac, A., Golub, K., Aalberg, T., Jatowt, A. (eds.) Digital Libraries for Open Knowledge. pp. 163–176. Springer International Publishing, Cham (2019)
13. Masanès, J.: Web archiving. Springer, Berlin, Germany (2006)
14. Ohio State University: Causal reasoning (2011), http://www.istarassessment.org/srdims/causal-reasoning-2/
15. Poursardar, F., Shipman, F.: How perceptions of web resource boundaries differ for institutional and personal archives. In: 2018 IEEE International Conference on Information Reuse and Integration (IRI). pp. 126–129 (2018). https://doi.org/https://doi.org/10.1109/IRI.2018.00026
16. QSR International: Nvivo product range (2016), http://www.qsrinternational.com/nvivo-product
17. Spaniol, M., Mazeika, A., Denev, D., Weikum, G.: "Catch me if you can": Visual analysis of coherence defects in web archiving. In: Proceedings of the 9th International Web Archiving Workshop (IWAW), Corfu, Greece, September 30 - October 1, 2009. pp. 27 – 37 (2009)